LA-UR-05-8420
*Approved for public release;
distribution is unlimited.*

Title:

A framework for assessing the impact of units of scholarly communication based on OAI-PMH harvesting of usage information

Author(s):

Johan Bollen
Herbert Van de Sompel

Submitted to:

CERN workshop on Innovations in Scholarly Communication (OAI4)
October 20-22, 2005
Geneva, Switzerland

A framework for assessing the impact of units of scholarly communication based on OAI-PMH harvesting of usage information.

Johan Bollen and Herbert van de Sompel

Digital Library Research & Prototyping Team

Research Library

Los Alamos National Laboratory

email: jbollen@lanl.gov

URL: http://public.lanl.gov/

“A cynic knows the price of everything, but the value of nothing”

- Evaluation of scholarly communication and production matters greatly:
  - Implications for individual researchers:
    - Publishing decisions
    - Career options
    - Collaboration choices
  - Resource allocation
    - Teams
    - Universities
    - Institutions
    - Nations
  - Bibliometrics
    - Trends
    - Process

- Science is unfortunately not evaluated on its inherent value, but by its “by-products”, in particular publication

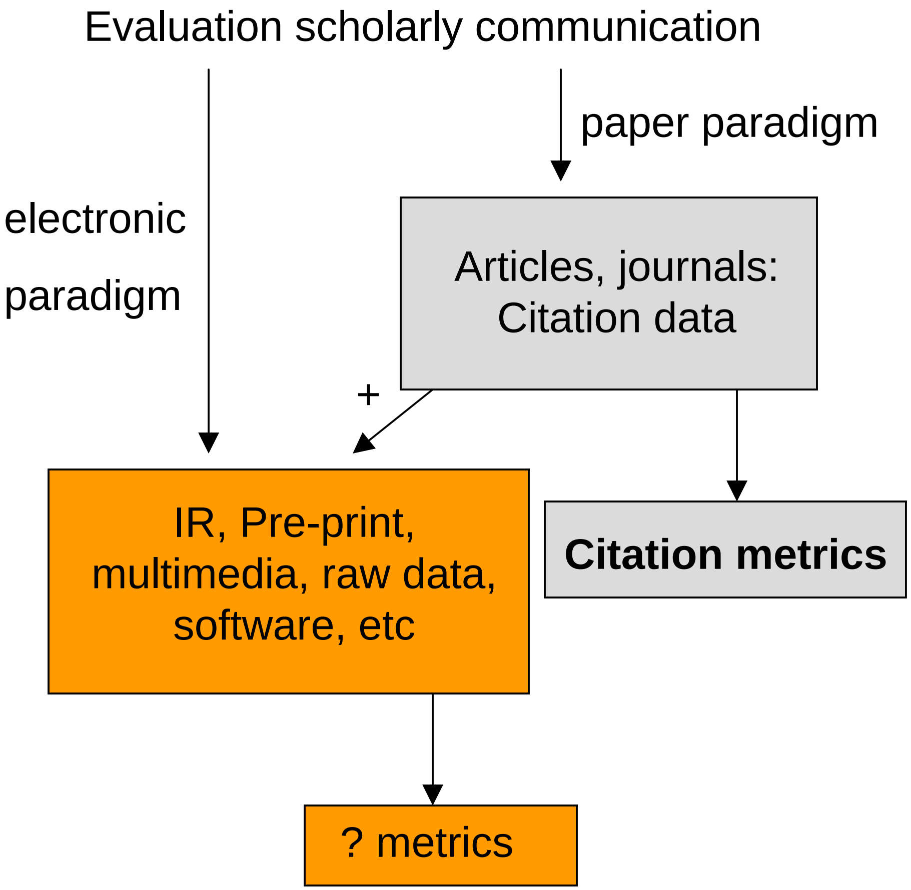
- A short-cut to expert evaluation: citation analysis
  - Citation indicates endorsement of published work by publishing peers
  - Count citations for researcher, departments, universities, nations to rank their prestige and productivity

- Pervasive approach in academia and elsewhere
  - Driven by availability of vetted citation data: Thomson Scientific ISI Journal Citation reports and Impact Factors
  - Applied in all areas of evaluation

What's going on in science evaluation?

- A paper economy
  - Citable, published literature only
  - Corresponding metrics: citation frequency
  - Someone dictates what we count and how we count

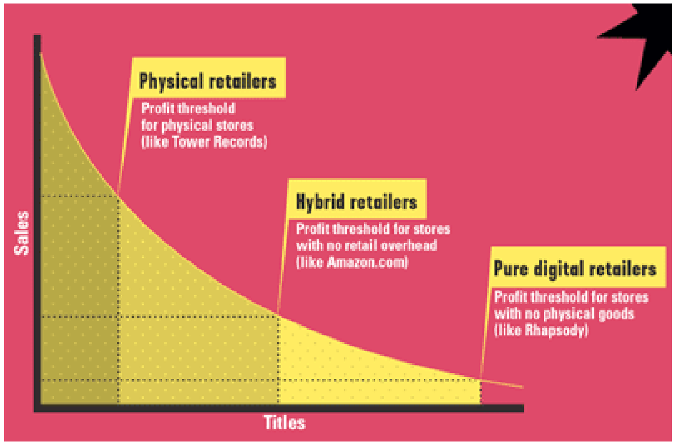
- Electronic paradigm
  - New models of communication
  - New models of scholarship
  - New metrics of evaluation?


Evaluation scholarly communication
paper paradigm
electronic
paradigm
Articles, journals: Citation data
+
IR, Pre-print, multimedia, raw data, software, etc
Citation metrics
? metrics

Science evaluation on the long tail

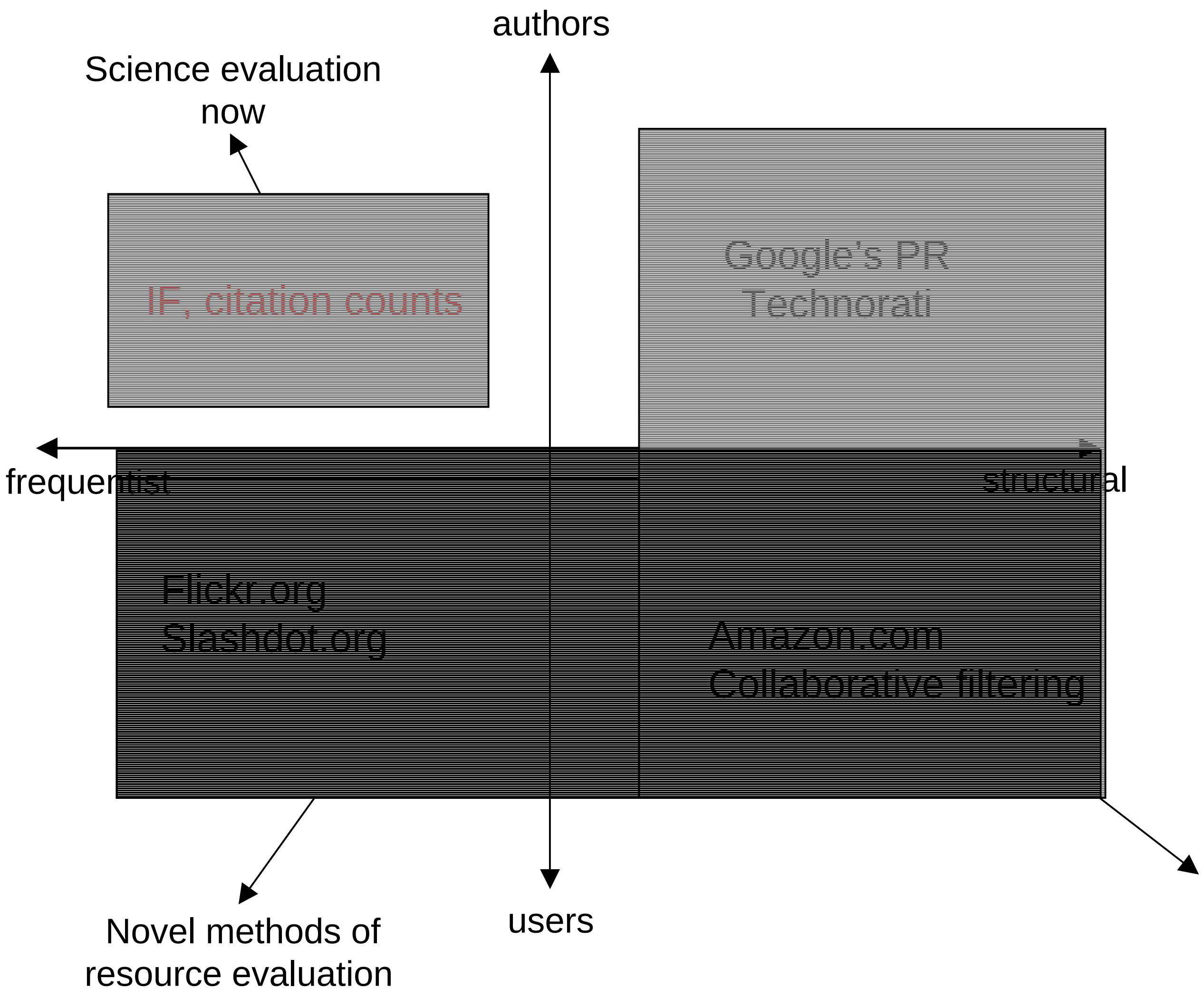
- Short front
  - Not everything can be published: selection and pruning process
  - Publication delays: the world as it was 3 years go
  - Use of citation as endorsement indicator: expert endorsement
  - Not every citation is counted: need for standardized, limited, and vetted data sets (ISI)

- **Long tail**
  - Everything will be published somehow
  - Immediate, electronic access to all stages of scholarly process
  - Many possible indicators of endorsement and interest (hyperlinks, readership, ratings)
  - No privileged access: all can count, all will be counted



Chris Anderson (2004),Wired, 12.10

The long tail: a user-driven revolution.



Evaluation of resources (quality, status, pretige) is required on all levels of our digital infrastructure:

•Most solutions adopted differ from citation analysis with regards to:

- •Different data sources
- •Different metrics

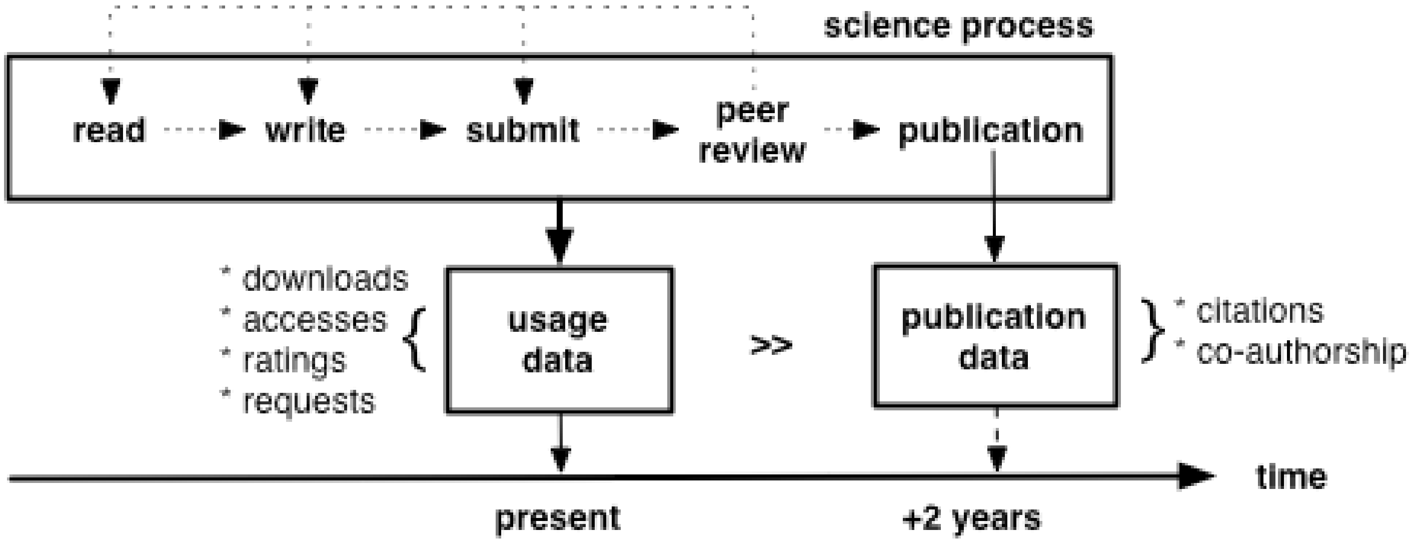
• Trend:

- •author controlled to user controlled
- •Frequentist to structural

LANL approach since 1999:
•Derive relational information from reader/user interest and citation data
* Structure defines status/prestige (social network science)

The importance of usage information.

- Recorded in the present (usage), not 3-4 years after fact (citation)
- Unlimited access, unlimited sample size
- Already recorded locally at many different information resources
- Reduced “social desirability bias”
- Recorded at all stages of the scholarly process
- Applies to all units of scholarly communication



Beyond linear click streams: mining usage data for item relationships

Citation

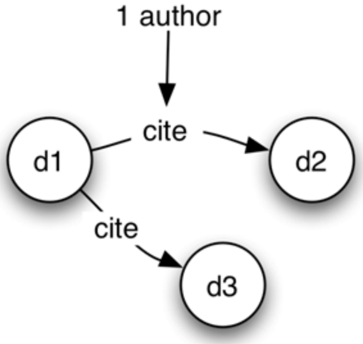
1) When an author cites B from A,
   A and B are related
2) The frequency of citation corresponds to
   degree of relatedness (journals)

↔

Clickstream/data mining approach:

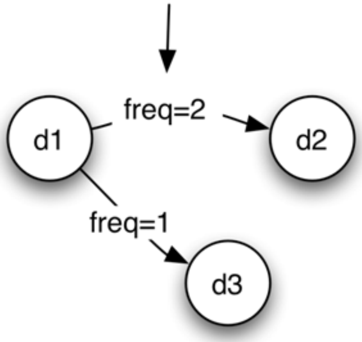
1) When a user downloads A and B,
   A and B may be related.
2) The co-download frequency corresponds to
   degree of relatedness (all docs)

citation data



usage: download log

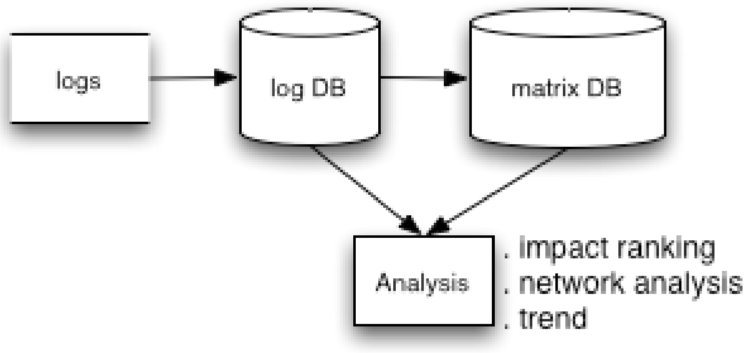
```
user1   date1   time1   d1 } co-download(d1,d2)
user1   date1   time2   d2                          } : freq(d1,d2)=2
user2   date2   time3   d1 } co-download(d1,d2)
user2   date2   time4   d2
user3   date3   time5   d1 } co-download(d1,d3)   : freq(d1,d3)=1
user3   date3   time6   d3
```



- *Kothari (2003). On using page cooccurence …*
- *Kim (2004). A clickstream-based collaborative…*
- *Sarwar (2001. Item-based collaborative filtering*

Local experiences at the LANL RL

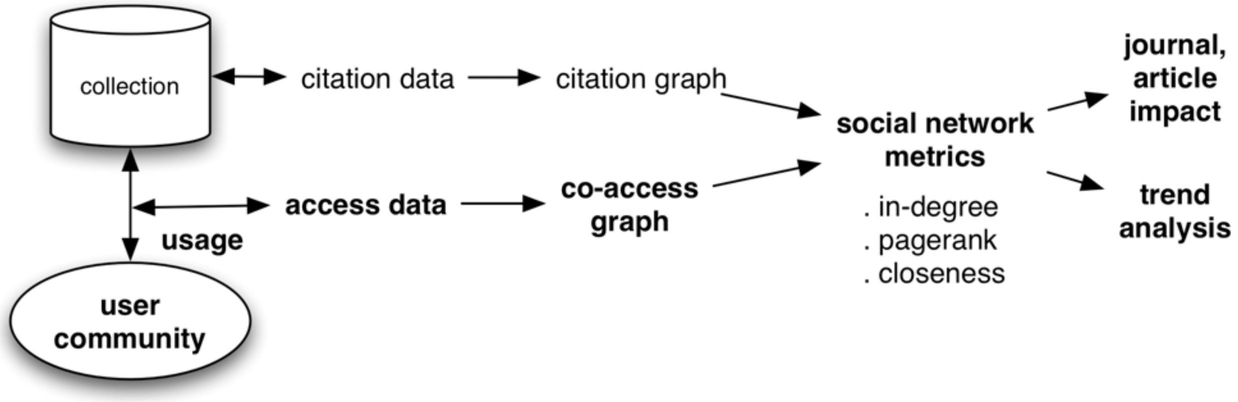
- Collect reliable usage data from multiple service providers at LANL
- Logs include all user expressions of interest
  - Request for metadata: author, abstract, reference, etc
  - Full-text downloads
  - …
- Most recent analysis: February 2004 to April 2005:
  - 392,455 usage events : any indication of preferences/interest
  - 5,866 users
  - 330,109 articles
  - 10,695 journals



Analysis focuses on:

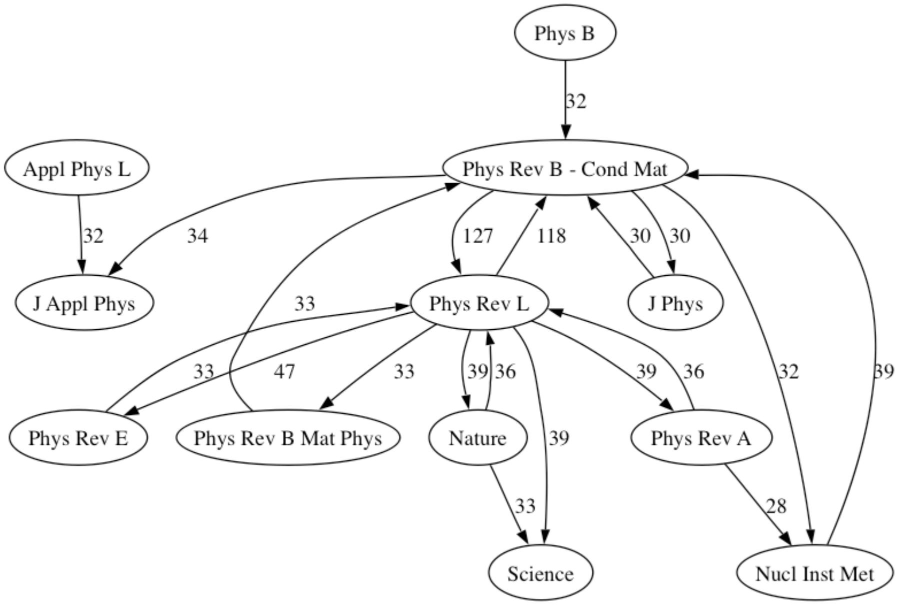
1) Journal impact metrics:
   - Frequency
   - In-degree (IF)
   - PageRank
2) Trend: comparison to citation data

LANL usage analysis methodology

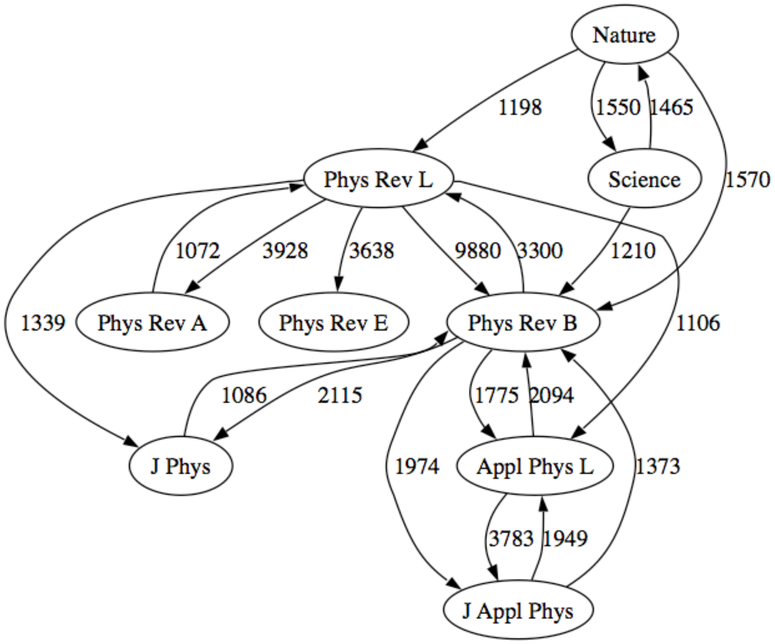


Usage: user activity that expresses interest or preference
Access data: particular instance(s) of usage (e.g. request abstract, download full-text)
Co-access: repeated instances of same user accessing pairs of items (documents)
Co-access graph: network of co-access data
Social network metrics: prestige from network structure

Journal matrix (02/2004-04/2005)



LANL 2004 data : top edges



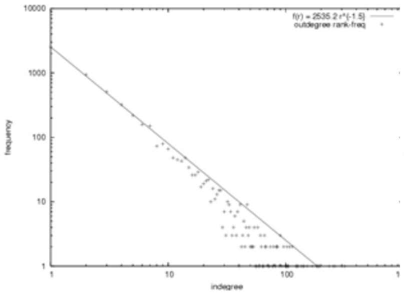
JCR2001 data : top edges

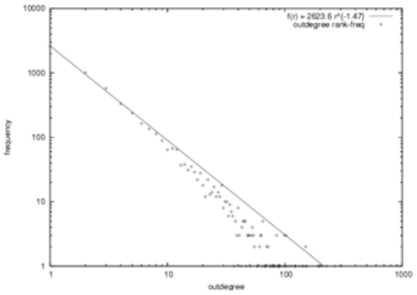
Resulting journal matrix:

- 33,256 edges
- 7,099 journals
- Density: 6.6 x 10^{-4}

Degree distribution:

- Power law, $f = k * r^{-a}$
- Small-world graphs



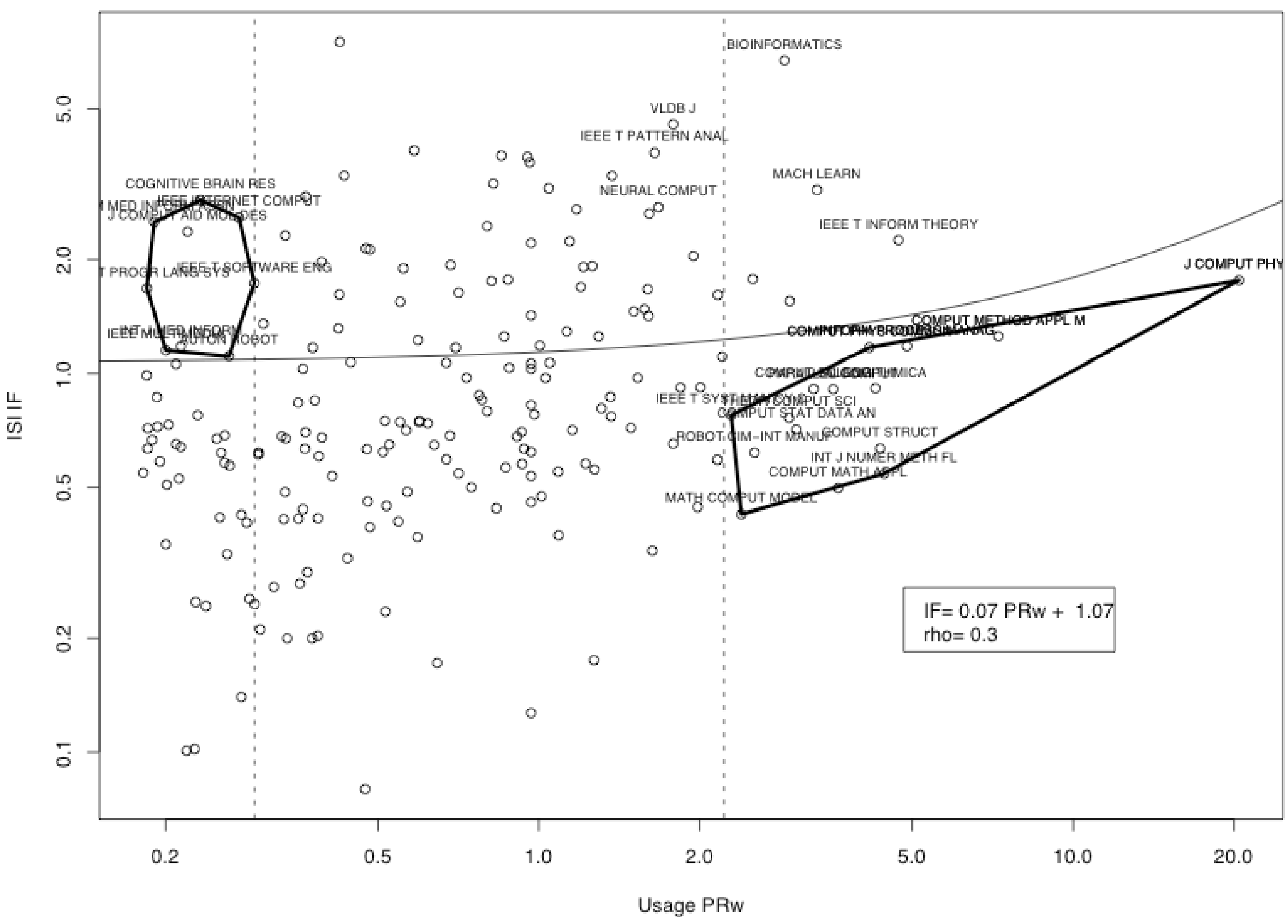


A comparison of LANL usage and citation impact

rank	Usage (PageRank)	IF (2003)	ISSN	Title (abbv.)
1	60.196	7.035	0031-9007	PHYS REV LETT
2	37.568	2.950	0021-9606	J CHEM PHYS
3	34.618	1.179	0022-3115	J NUCL MATER
4	31.132	2.202	1063-651X	PHYS REV E
5	30.441	2.171	0021-8979	J APPL PHYS
6	30.128	30.979	0028-0836	NATURE
7	29.972	29.781	0036-8075	SCIENCE
8	27.187	6.516	0002-7863	J AM CHEM SOC
9	24.602	4.049	0003-6951	APPL PHYS LETT
10	23.631	2.992	0148-0227	J GEOPHYS RES

Users and authors agree. Somewhat.

LANL04 Usage Weighted PageRank vs. 2003 IF (Computer Science)

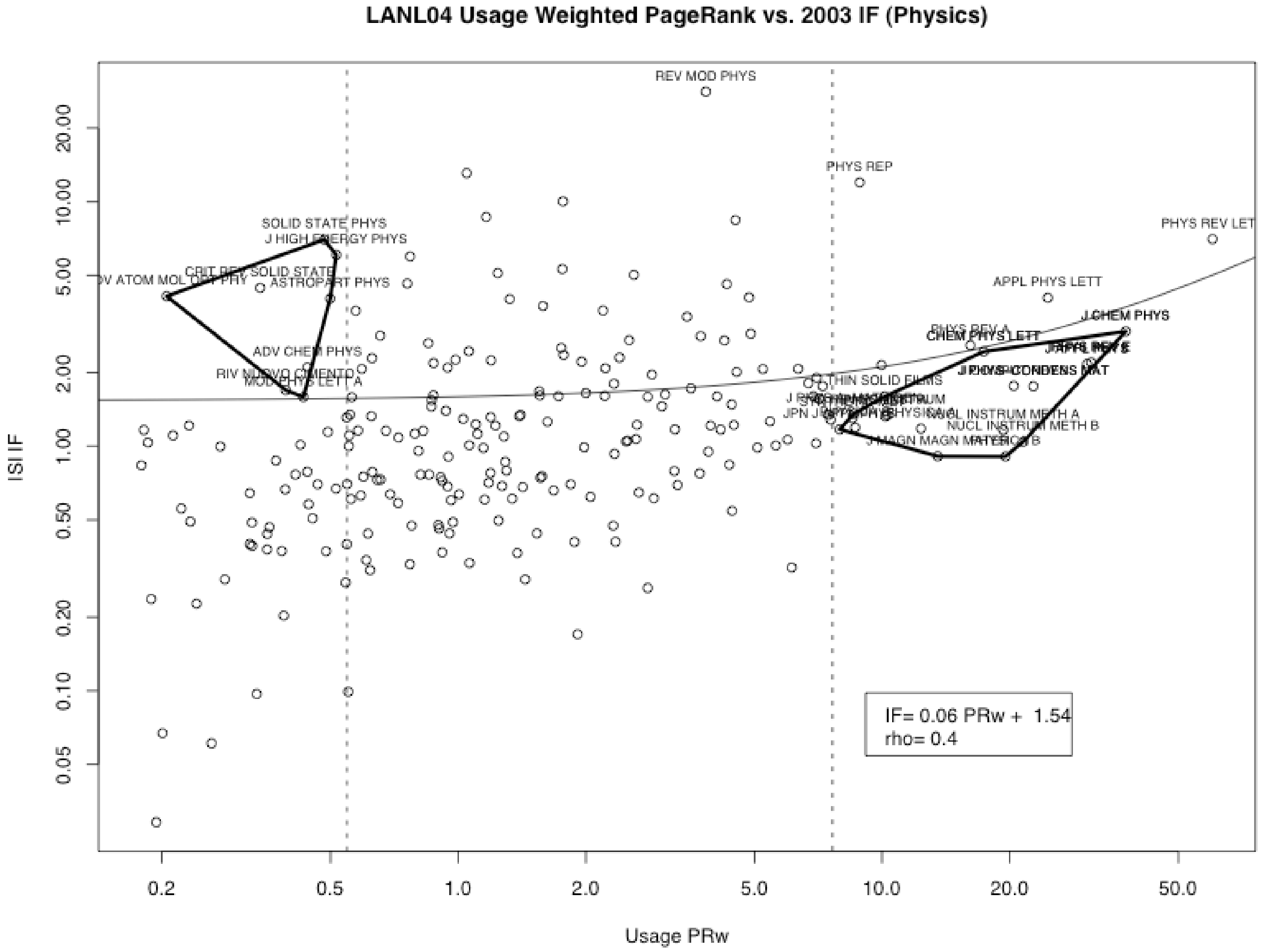


Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ OAI4 - Geneva, Switzerland

RESEARCH
LIBRARY

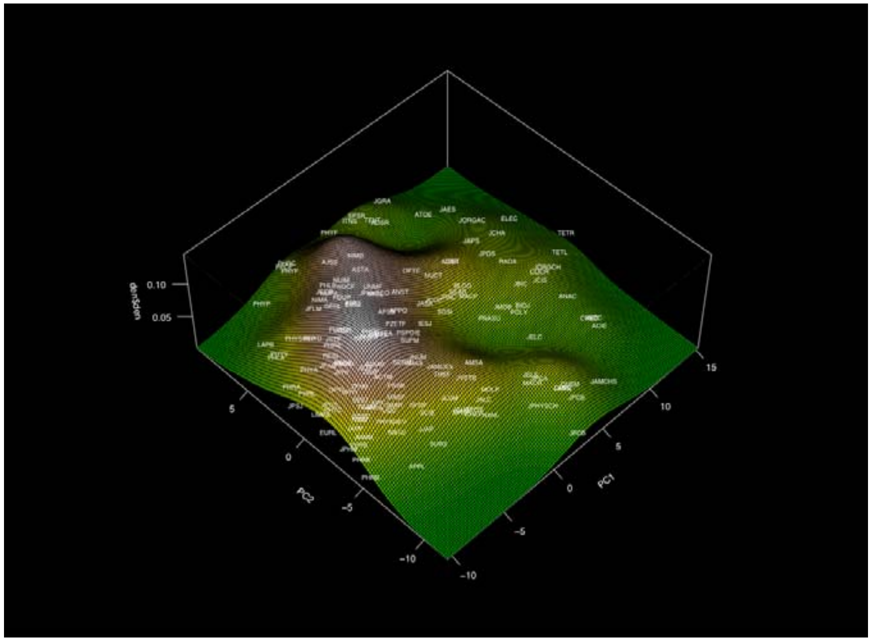
Los Alamos
NATIONAL LABORATORY

Users and authors agree. Somewhat.

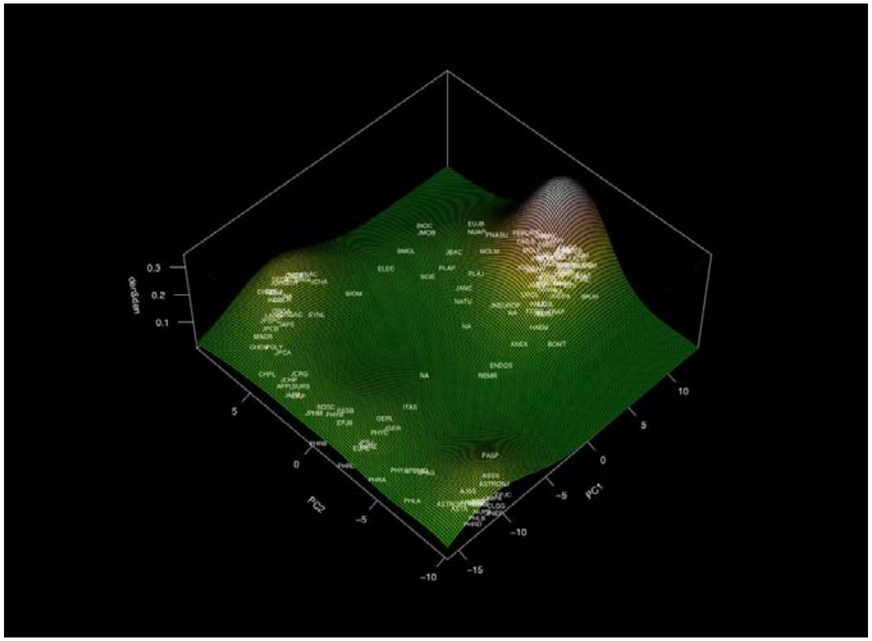


LANL04 Usage Weighted PageRank vs. 2003 IF (Physics)

Information landscapes:
studying the structure and evolution of science



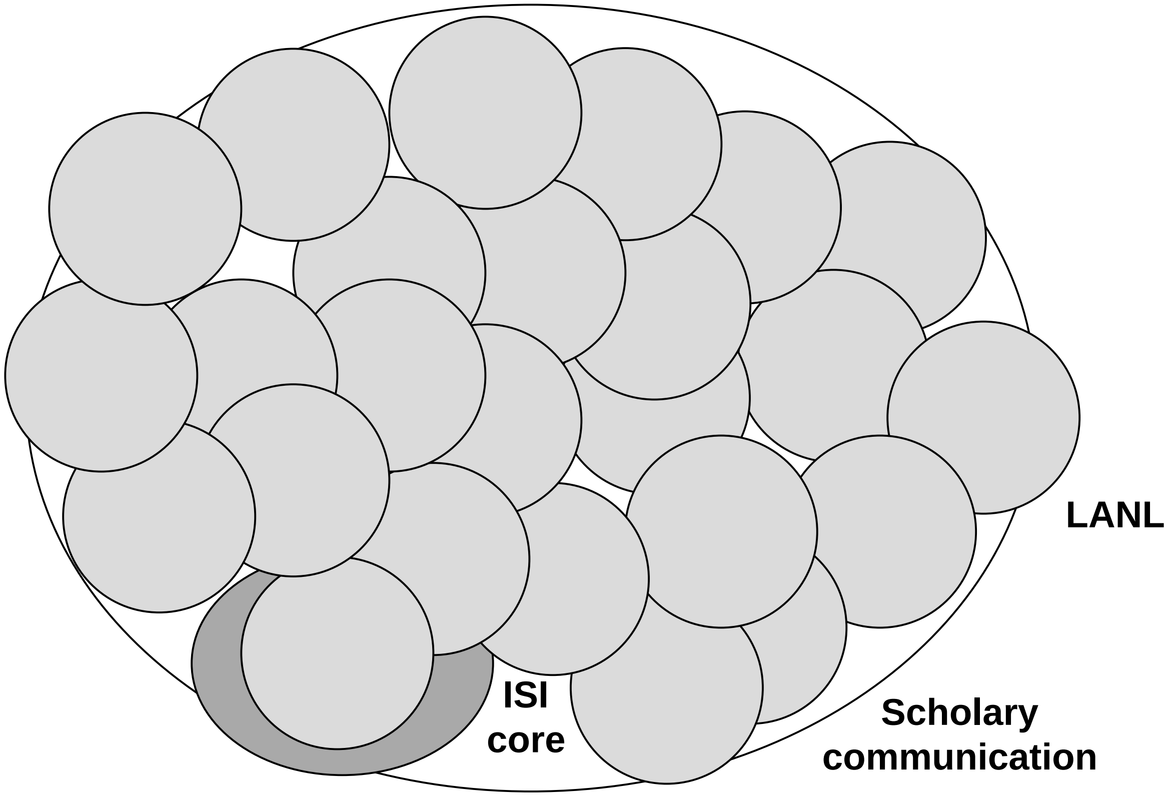
LANL04



JCR03

- **Two component model**
- **PC1: Life vs. natural science**
- **PC2: Microscopic vs. macroscopic**
- **Z-axis: cluster density**

From local usage to global coverage



- Local usage is interesting
  - Informs local collection management
  - Prominent communities can inform assessments of science trends
  - Covers wide range of communication items
  - Immediate availability
- Global, aggregated usage data is even more interesting
  - Monitor science as it takes place
  - Replace/augment/validate proprietary data sets
  - Allow free-form aggregation:
    - Clusters of institutions
    - Focus on sub-domains and communities

Challenges applying usage data in “global” scholarly evaluation .

- Institutional rights and biases
  - Registered locally:
    - Proprietary?
    - Privacy issues
- Standardization
  - What usage is being recorded?
  - How is it registered and stored?
- Aggregation and scalability
  - Joining logs from different origins is required for “global” analysis beyond institution
  - Standardization and scalability issues
- Metrics
  - Frequentist metrics indicate popularity not impact/quality
  - Structual metrics require structural data: linear usage logs?

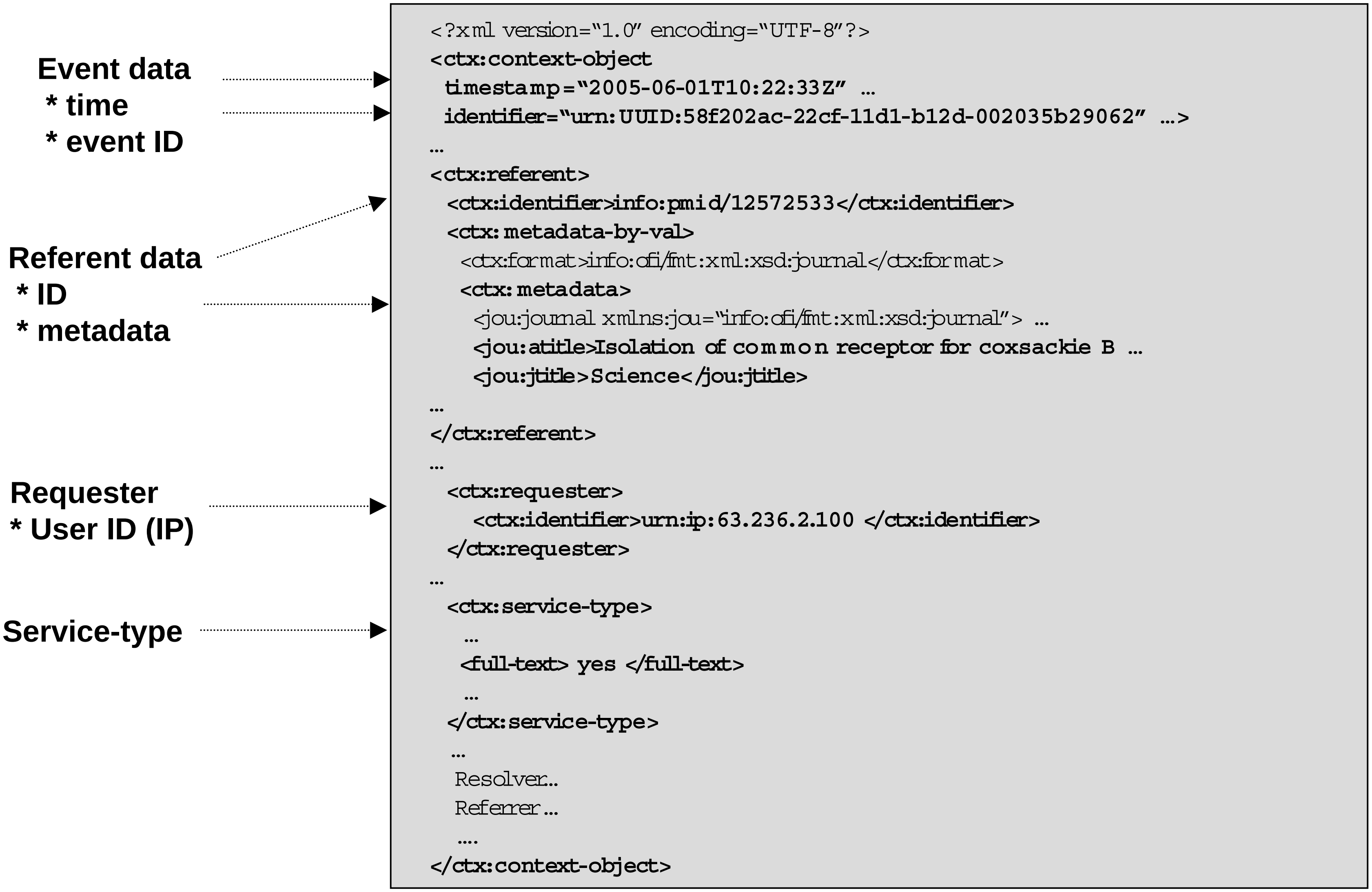
- LANL solution (3 components):
  - Standards: OAI-PMH and OpenURL ContextObjects for log harvesting
    - Logs exposed and harvested using OAI-PMH
    - Usage data represented using OpenURL ContextObjects
  - Data mining:
    - Derive document relationships from access sequences
  - Metrics:
    - Recommender system
    - Structural metrics of impact/prestige/prominence

Transport and representation

- OAI-PMH
  - Data provider exposes access logs via OAI-PMH repository
  - Metadata exposed = usage event data
    - Who?
    - What?
    - When?
    - How?
  - Expressed as XML OpenURL ContextObjects
  - Harvested by aggregator

- **OpenURL ContextObjects for log data**:
  - Each object assigned globally unique ID, i.e. UUID
  - Timestamp: recorded time of event
- **Referent**:
  - One or more identifiers (URIs) for resource involved in event
  - By_value Metadata Descriptor
- **Requeste**r:
  - Requester associated with event, i.e. user
  - Currently only the IP address of the Requester's machine, urn:ip:....
- **Service-type**: By_value Metadata Descriptor for the Service Types involved in the event
- **Resolve**r: One or more identifiers for the OpenURL Resolver
- **Referrer**: identifiers for referrer involved in event

OpenURL ContextObject for usage event data



Event data
*** time**
*** event ID**

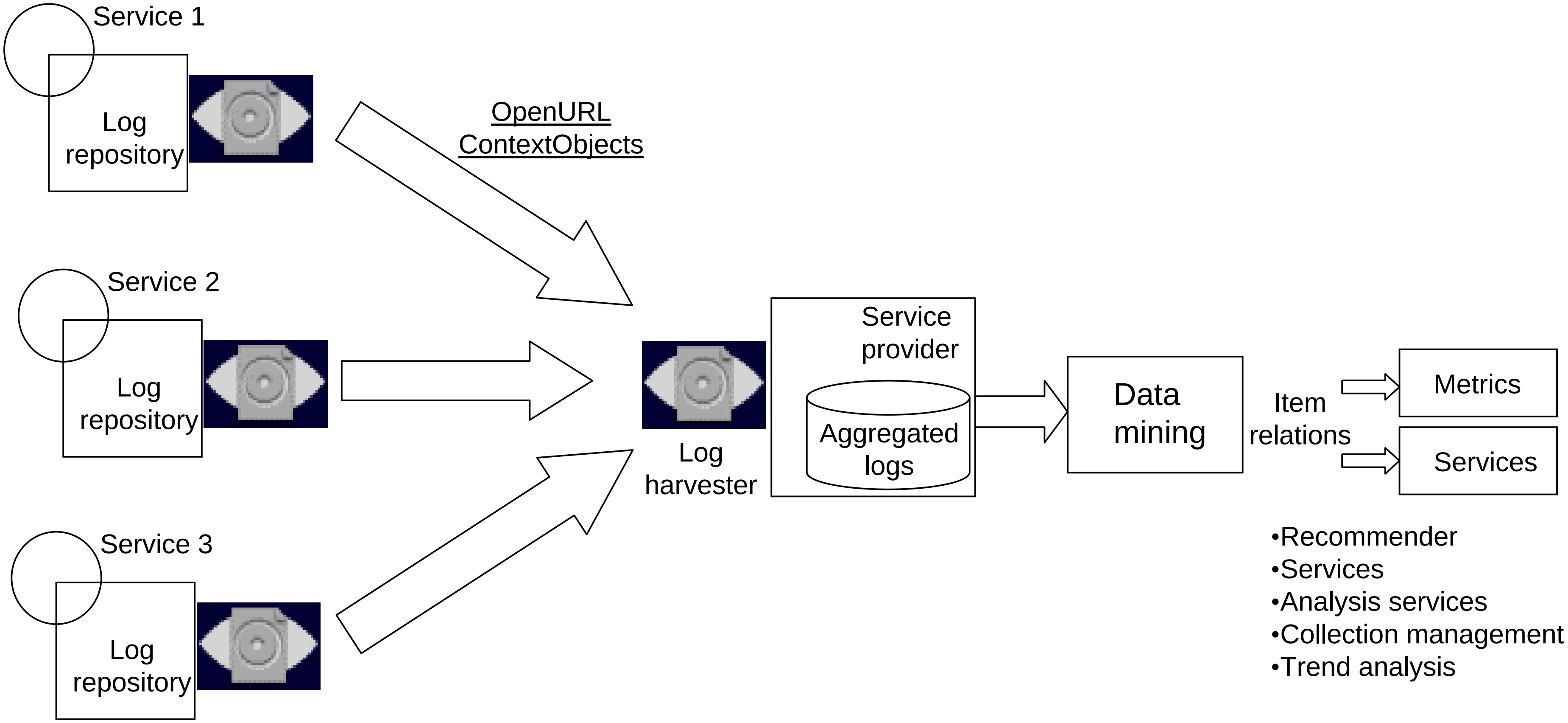
Referent data
*** ID**
*** metadata**

Requester
*** User ID (IP)**

Service-type

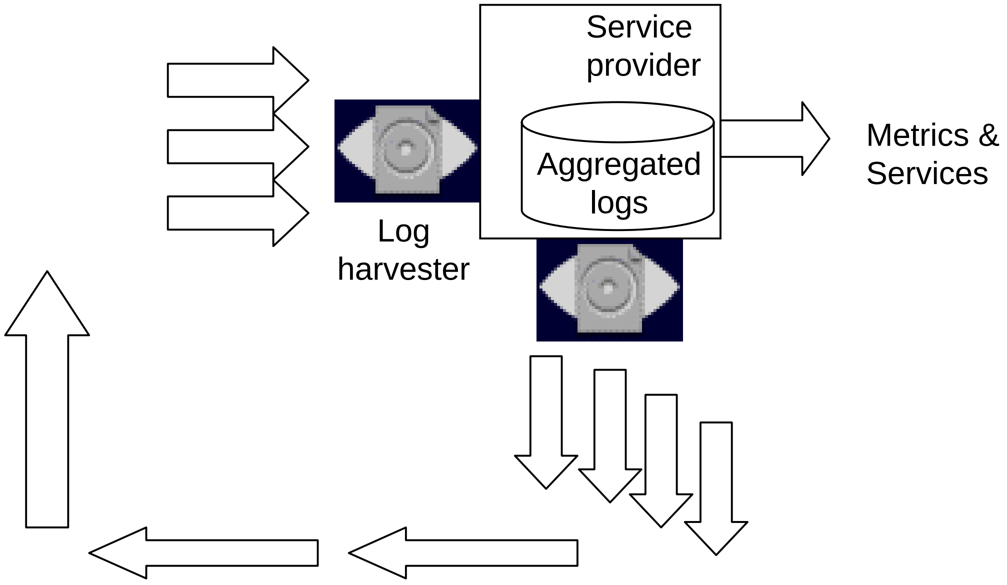
```
<?xml version="1.0" encoding="UTF-8"?>
<ctx:context-object
 timestamp="2005-06-01T10:22:33Z" …
 identifier="urn:UUID:58f202ac-22cf-11d1-b12d-002035b29062" …>
…
<ctx:referent>
  <ctx:identifier>info:pmid/12572533</ctx:identifier>
  <ctx: metadata-by-val>
   <ctx:format>info:ofi/fmt:xml:xsd:journal</ctx:format>
   <ctx: metadata>
    <jou:journal xmlns:jou="info:ofi/fmt:xml:xsd:journal"> …
    <jou:atitle>Isolation of common receptor for coxsackie B …
    <jou:jtitle>Science</jou:jtitle>
…
</ctx:referent>
…
  <ctx:requester>
    <ctx:identifier>urn:ip:63.236.2.100 </ctx:identifier>
  </ctx:requester>
…
  <ctx:service-type>
    …
   <full-text> yes </full-text>
    …
  </ctx:service-type>
   …
   Resolver…
   Referrer …
   ….
</ctx:context-object>
```

General architecture



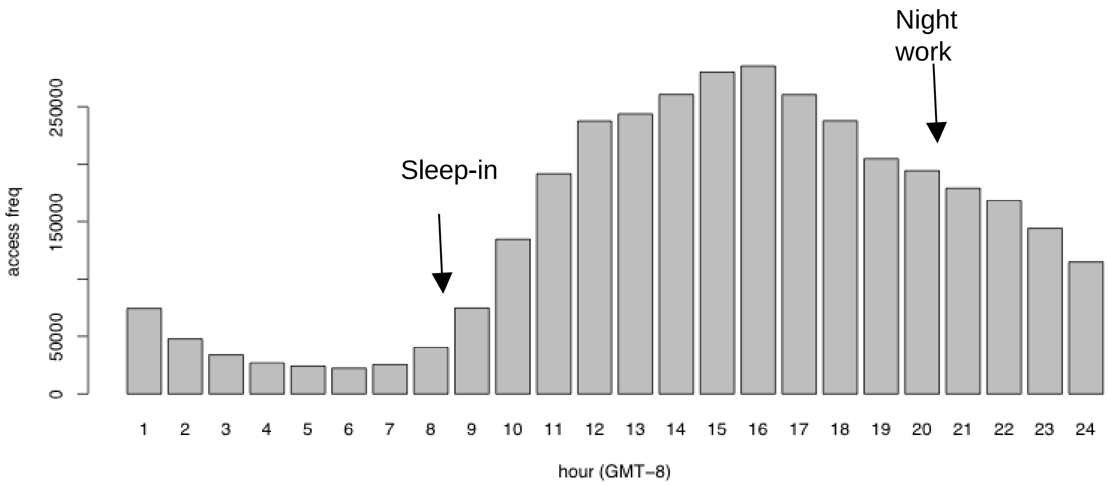
- Recommender
- Services
- Analysis services
- Collection management
- Trend analysis

Community-centric vs. global services: incentives to federate



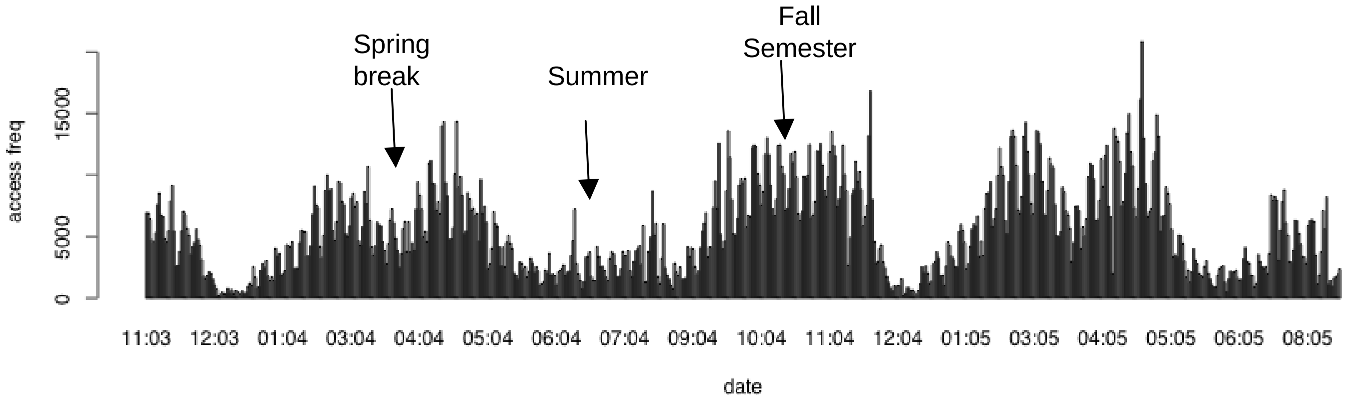
- Every instance can act as both a aggregator and data provider: re-expose or not?
- Local Benefits: Benefits when not exposing log data:
  - Local recommender services
  - Local collection management data
  - Local trend data
- But wouldn't you want more?
- Exposing logs for harvesting:
  - Obtain permission to harvest from global or federated data set
  - Acquire global services
  - Be represented: is your community included?
- Possible mergence of 3rd federators
  - Provides trusted global analysis
  - Services based on global data
  - Re-exposes vetted, federated data

Some statistics: the academic rhythm


Night work

Sleep-in

- Logs collected at 9 institutions and LANL federated
- 3,507,484 unique events
- 2,133,556 unique documents
-167,204 unique agents (users)
- Recorded: November 2003 to August 2005
- 67% article events, 25% journal events


Spring break

Summer

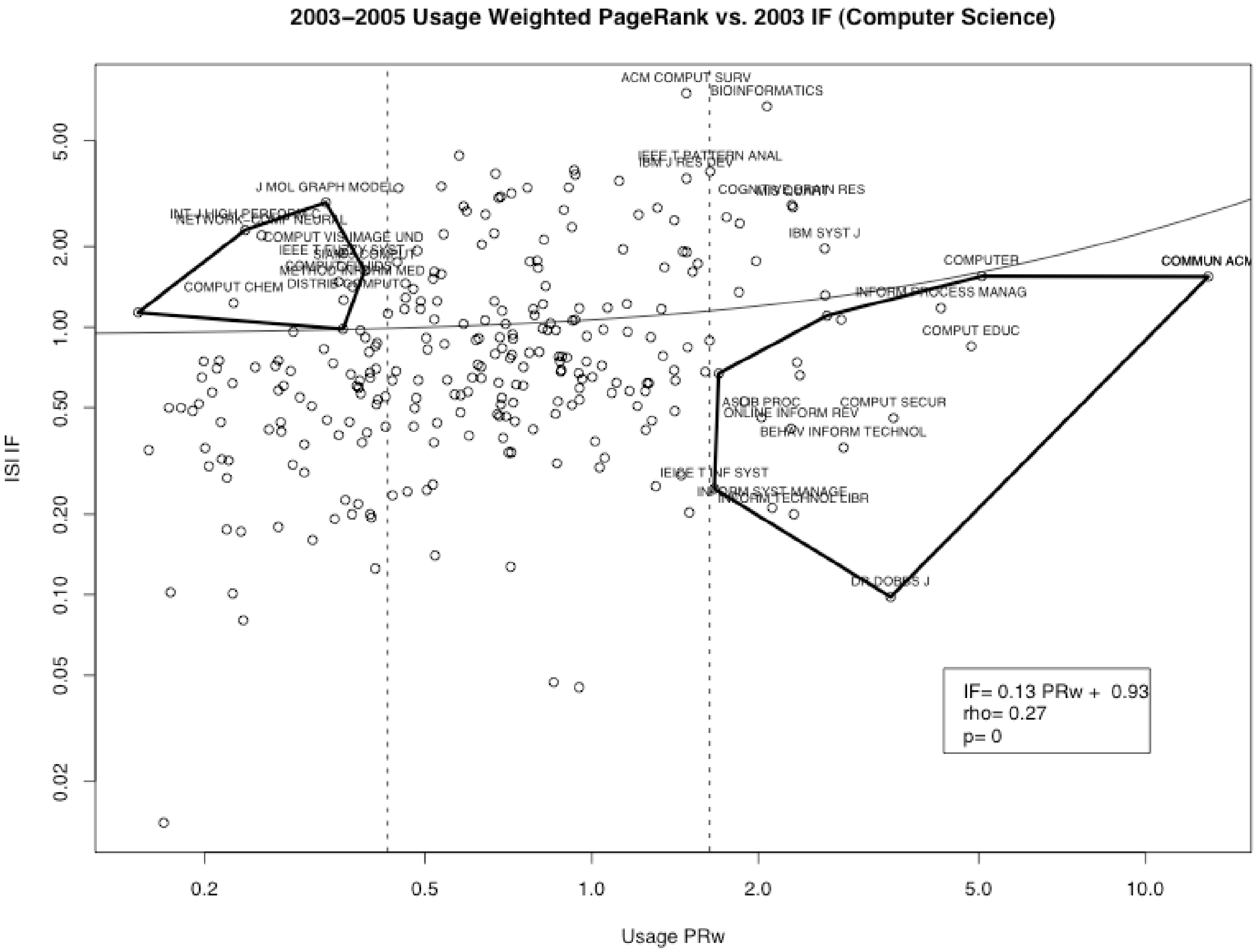
Fall Semester

QuickTime™ and a
Graphics decompressor
are needed to see this picture.

Results: journal ranking

rank	Usage (PageRank)	IF (2003)	ISSN	Title (abbv.)
1	78.565	21.455	0098-7484	JAMA-J AM MED ASSOC
2	71.414	29.781	0036-8075	SCIENCE
3	60.373	30.979	0028-0836	NATURE
4	40.828	3.779	0890-8567	J AM ACAD CHILD PSY
5	39.708	7.157	0002-953X	AM J PSYCHIAT
6	38.113	34.833	0028-4793	NEW ENGL J MED
7	37.492	3.363	0090-0036	AM J PUBLIC HEALTH
8	37.031	2.591	0195-9131	MED SCI SPORT EXER
9	27.248	0.998	0309-2402	J ADV NURS
10	26.987	5.692	0002-9165	AM J CLIN NUTR

Comparison journal usage and citation IF

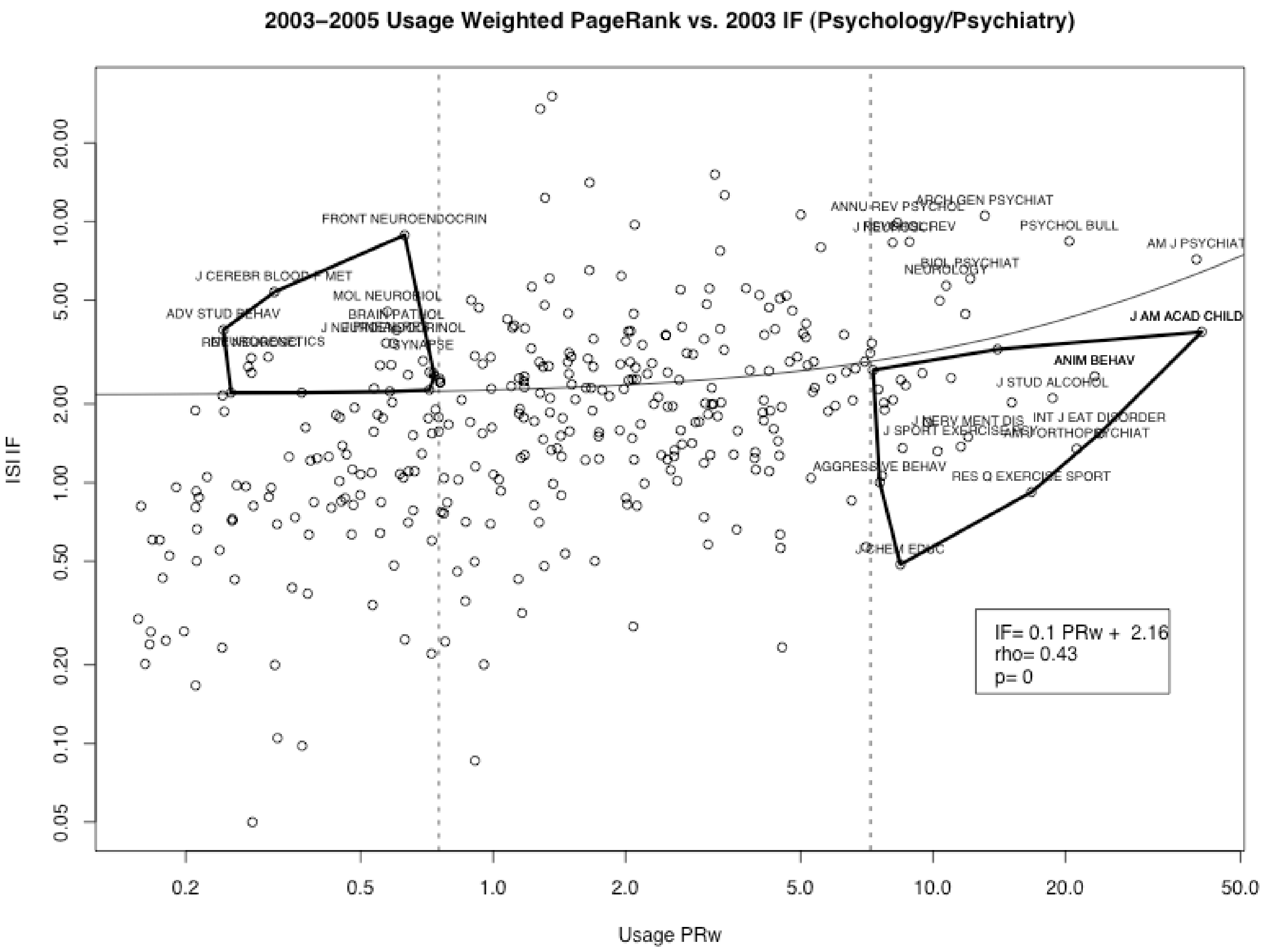


RESEARCH LIBRARY

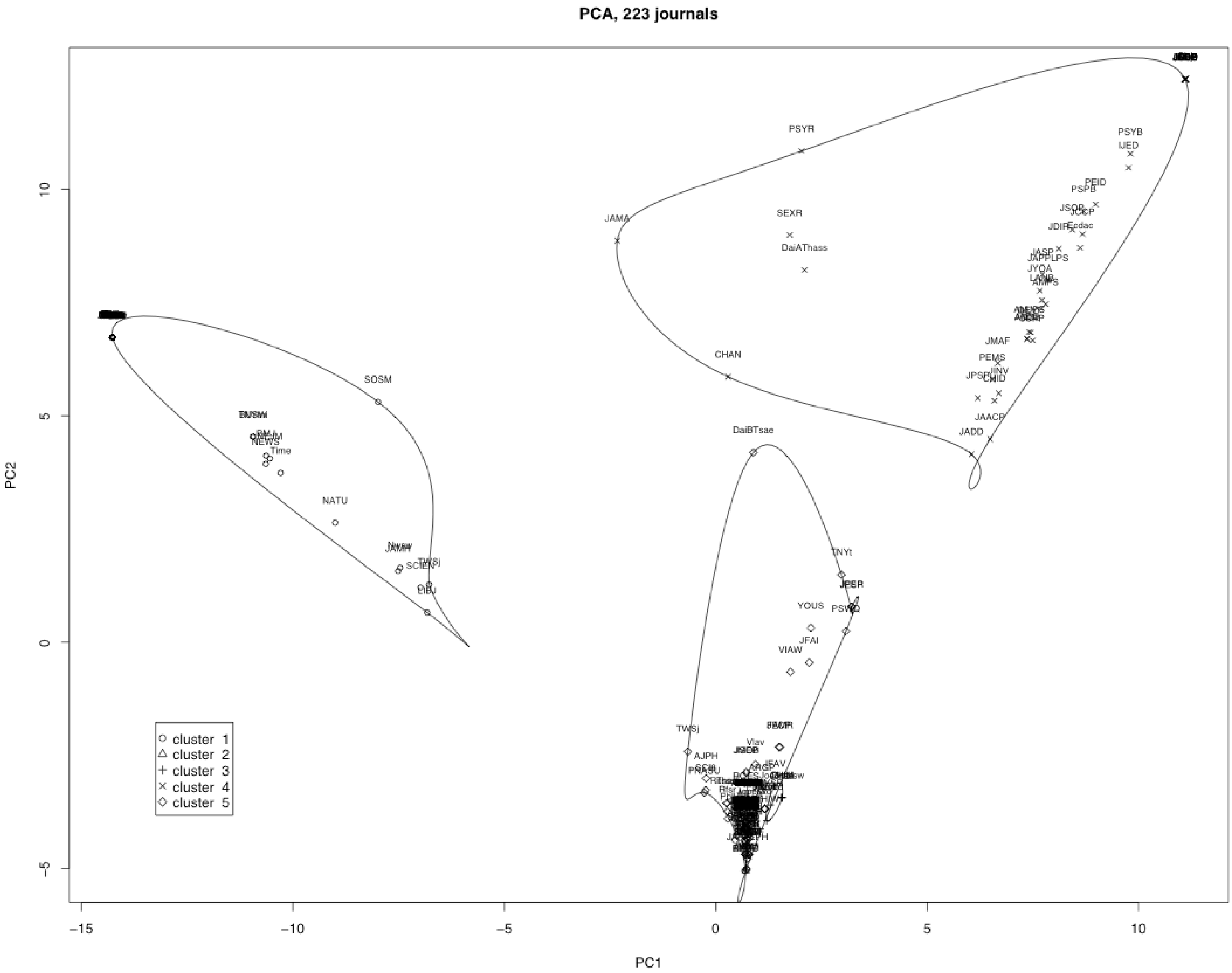
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ OAI4 - Geneva, Switzerland

Los Alamos NATIONAL LABORATORY

Comparison journal usage and citation IF, contd.



Mapping the structure of science



Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ OAI4 - Geneva, Switzerland

RESEARCH
LIBRARY

Los Alamos
NATIONAL LABORATORY

Article level data

Frequency	Article
441	Cardell Jacobson (2001) Religion, Religiosity, and Attribution of Responsibility. Research in the social scientific study of religion 12: 117
273	J. Sloan (2004) Respondent Misreporting of Drug Use in Self-Reports: Social Desirability and Other Correlates. Journal of Drug Issues 34:269
242	Zaborski, E. R. (2002). Observations on feeding behavior by the terrestrial flatworm Bipalium adventitium. Am. Midl. Nat. 148:201
187	Genz (1998) Working the reference desk. Library Trends, 46:505
185	C. D. Fiore (1998) The numbers game: how to fatten your budget by using statistics", School Library Journal, 44(3)
163	Van Horn (2002). The Digital Millennium Copyright Act and Other Egregious Laws. Phi Delta Kappan. 84:248
185	Iglehart (2002).: Hispanic and African American Youth: Life after Foster Care Emancipation. Journal of ethnic & cultural diversity in social work. 11:79
145	Stoffle (1994), “No Place for Neutrality: the Case for Multiculturalism,” *Library Journal* 119:46
133	Simpson (1999) Managing Copyright in Schools. Knowledge quest. 28:18
129	Heppermann (1998). Little house on the bottom line. The horn book magazine 74:689

Usage-based recommender system

- Operates on network derived from aggregated usage
- Starts from (set of) documents (articles or journals)
- Scans usage network links for direct and indirectly related documents
- Results:
  - Scalable
  - Highly efficient
  - Highly relevant results derived from accumulated, aggregated usage

Movie: article level recommendations

Movie: journal level recommendations

Conclusion

- Scholarly communication is going through a revolution
- Scholarly evaluation will too! Focus will be on
  - Immediacy
  - Representativeness
  - Openness, standards and scalability
  - Acknowledging structural aspects of prestige and impact in the scholarly community
- User driven evaluation offers an interesting alternative to current short-front evaluation methods in a long-tail world

- Feasibility of usage analysis demonstrated at local and semi-global level
  - LANL results indicate:
    - Possibility of local prestige and impact ranking
    - Additional usage-based services such as recommender systems possible
  - Aggregated data and analysis:
    - Large-scale aggregation demonstrated scalability
    - Use of existing standards ensures openness, ability of all to participate
    - Possibility of spontaneous emergence of vetting and standardization system for usage quality indicators
    - Enticing community and global recommender services offer further incentives to adopt locally and collaborate globally

Some papers:

- **J. Bollen, H. V. de Sompel, J. Smith, and R. Luce**. Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419-1440, 2005.
- **J. Bollen, R. Luce, S. Vemulapalli, and W. Xu**. Detecting research trends in digital library readership. In *Proceedings of the Seventh European Conference on Digital Libraries (LNCS 2769)*, pages 24-28, Trondheim, Norway, August 18 2003. Springer-Verlag.
- **J. Bollen, R. Luce, S. Vemulapalli, and W. Xu**. Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*, 9(5), 2003.